

Adaptive Estimation of Shannon Entropy

Yanjun Han, *Student Member, IEEE*, Jiantao Jiao, *Student Member, IEEE*, and Tsachy Weissman, *Fellow, IEEE*

Abstract

We consider estimating the Shannon entropy of a discrete distribution P from n i.i.d. samples. Recently, Jiao, Venkat, Han, and Weissman, and Wu and Yang constructed approximation theoretic estimators that achieve the minimax L_2 rates in estimating entropy. Their estimators are consistent given $n \gg \frac{S}{\ln S}$ samples, where S is the alphabet size, and it is the best possible sample complexity. In contrast, the Maximum Likelihood Estimator (MLE), which is the empirical entropy, requires $n \gg S$ samples.

In the present paper we significantly refine the minimax results of existing work. To alleviate the pessimism of minimaxity, we adopt the adaptive estimation framework, and show that the minimax rate-optimal estimator in Jiao, Venkat, Han, and Weissman achieves the minimax rates simultaneously over a nested sequence of subsets of distributions P , without knowing the alphabet size S or which subset P lies in. In other words, their estimator is adaptive with respect to this nested sequence of the parameter space, which is characterized by the entropy of the distribution. We also characterize the maximum risk of the MLE over this nested sequence, and show, for every subset in the sequence, that the performance of the minimax rate-optimal estimator with n samples is essentially that of the MLE with $n \ln n$ samples, thereby further substantiating the generality of the phenomenon discovered by Jiao, Venkat, Han, and Weissman.

Index Terms

adaptive estimation, entropy estimation, best polynomial approximation, high dimensional statistics, large alphabet, minimax optimality

I. INTRODUCTION

Shannon entropy $H(P)$, defined as

$$H(P) \triangleq \sum_{i=1}^S p_i \ln \frac{1}{p_i}, \quad (1)$$

is one of the most fundamental quantities of information theory and statistics, which emerged in Shannon's 1948 masterpiece [1] as the answer to foundational questions of compression and communication.

Consider the problem of estimating Shannon entropy $H(P)$ from n i.i.d. samples. Classical theory is mainly concerned with the case where the number of samples $n \rightarrow \infty$, while the alphabet size S is fixed. In that scenario, the maximum likelihood estimator (MLE), $H(P_n)$, which plugs in the empirical distribution into the definition of entropy, is *asymptotically efficient* [2, Thm. 8.11, Lemma 8.14] in the sense of the Hájek convolution theorem [3] and the Hájek–Le Cam local asymptotic minimax theorem [4]. It is therefore not surprising to encounter the following quote from the introduction of Wyner and Foster [5] who considered entropy estimation:

“The plug-in estimate is universal and optimal not only for finite alphabet i.i.d. sources but also for finite alphabet, finite memory sources. On the other hand, practically as well as theoretically, these problems are of little interest.”

In contrast, various modern data-analytic applications deal with datasets which do not fall into the regime of fixed alphabet and $n \rightarrow \infty$. In fact, in many applications the alphabet size S is comparable to, or even larger than the number of samples n . For example:

- Corpus linguistics: about half of the words in the Shakespearean canon appeared only once [6].
- Network traffic analysis: many customers or website users are seen a small number of times [7].
- Analyzing neural spike trains: natural stimuli generate neural responses of high timing precision resulting in a massive space of meaningful responses [8]–[10].

A. Existing literature

The problem of entropy estimation in the large alphabet regime (or non-asymptotic analysis) has been investigated extensively in various disciplines, which we refer to [11] for a detailed review. One recent breakthrough in this direction came from Valiant and Valiant [12], who constructed the first explicit entropy estimator whose sample complexity is $n \asymp \frac{S}{\ln S}$ samples, which they also proved to be necessary. It was also shown in [13] [14] that the MLE requires $n \asymp S$ samples, implying that MLE is strictly sub-optimal in terms of sample complexity.

However, the aforementioned estimators have not been shown to achieve the minimax L_2 rates. In light of this, Jiao et al. [11], and Wu and Yang in [15] independently developed schemes based on approximation theory, and obtained the minimax L_2 convergence rates for the entropy. Furthermore, Jiao et al. [11] proposed a general methodology for estimating functionals, and showed that for a wide class of functionals (including entropy, mutual information, and Rényi entropy), their methodology

can construct minimax rate-optimal estimators whose performance with n samples is essentially that of the MLE with $n \ln n$ samples. They also obtained minimax L_2 rates for estimating a large class of functionals. On the practical side, Jiao et al. [16] showed that the minimax rate-optimal estimators introduced in [11] can lead to consistent and substantial performance boosts in various machine learning algorithms.

Recall that the minimax risk of estimating functional $F(P)$ is defined via $\inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{F} - F(P) \right)^2$, where \mathcal{M}_S denotes all distributions with alphabet size S , and the infimum is taken with respect to all estimators \hat{F} . Correspondingly, the maximum risk of MLE $F(P_n)$, which evaluates the functional $F(\cdot)$ at the empirical distribution P_n , is defined via $\sup_{P \in \mathcal{M}_S} \mathbb{E}_P (F(P_n) - F(P))^2$. The following table in Jiao et al. [11] summaries the minimax L_2 rates and the L_2 rates of MLE in estimating $H(P)$ and $F_\alpha(P) \triangleq \sum_{i=1}^S p_i^\alpha$. Whenever there are two terms, the first term corresponds to squared bias, and the second term corresponds to variance. It is evident that one can obtain the minimax rates from the L_2 rates of MLE via replacing n with $n \ln n$ in the dominating (bias) terms. We adopt the following notation: $a_n \preceq b_n$ means $\sup_n a_n/b_n < \infty$, $a_n \succeq b_n$ means $b_n \preceq a_n$, $a_n \asymp b_n$ means $a_n \preceq b_n$ and $a_n \succeq b_n$, or equivalently, there exists two universal constants c, C such that

$$0 < c < \liminf_{n \rightarrow \infty} \frac{a_n}{b_n} \leq \limsup_{n \rightarrow \infty} \frac{a_n}{b_n} < C < \infty. \quad (2)$$

	Minimax L_2 rates	L_2 rates of MLE
$H(P)$	$\frac{S^2}{(n \ln n)^2} + \frac{\ln^2 S}{n} \quad (n \succeq \frac{S}{\ln S}) \text{ ([11], [15])}$	$\frac{S^2}{n^2} + \frac{\ln^2 S}{n} \quad (n \succeq S) \text{ [14]}$
$F_\alpha(P), 0 < \alpha \leq \frac{1}{2}$	$\frac{S^2}{(n \ln n)^{2\alpha}} \quad (n \succeq S^{1/\alpha} / \ln S, \ln n \preceq \ln S) \text{ ([11])}$	$\frac{S^2}{n^{2\alpha}} \quad (n \succeq S^{1/\alpha}) \text{ [14]}$
$F_\alpha(P), \frac{1}{2} < \alpha < 1$	$\frac{S^2}{(n \ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n} \quad (n \succeq S^{1/\alpha} / \ln S) \text{ ([11])}$	$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n} \quad (n \succeq S^{1/\alpha}) \text{ [14]}$
$F_\alpha(P), 1 < \alpha < \frac{3}{2}$	$(n \ln n)^{-2(\alpha-1)} \quad (S \succeq n \ln n) \text{ ([11])}$	$n^{-2(\alpha-1)} \quad (S \succeq n) \text{ [14]}$
$F_\alpha(P), \alpha \geq \frac{3}{2}$	$n^{-1} \text{ [14]}$	n^{-1}

TABLE I: Comparison of the minimax L_2 rates and the L_2 rates of MLE in estimating $H(P)$ and $F_\alpha(P) \triangleq \sum_{i=1}^S p_i^\alpha$. Whenever there are two terms, the first term corresponds to squared bias, and the second term corresponds to variance. It is evident that one can obtain the minimax rates from the L_2 rates of MLE via replacing n with $n \ln n$ in the dominating (bias) terms.

B. Refined minimaxity: adaptive estimation

One concern the readers may have about results on minimax rates is that they are too pessimistic. Indeed, in the definition $\inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{F} - F(P) \right)^2$, we have considered the worst case distribution P over all possible distributions supported on S elements, and it would be disappointing if the estimator in Jiao et al. [11] turned out to behave sub-optimally when we consider distributions lying in subsets of \mathcal{M}_S . A usual approach to alleviate this concern is the adaptive estimation framework, which we briefly review below.

The primary approach to alleviate the pessimism of minimaxity in statistics is the construction of adaptive procedures, which has gained particular prominence in nonparametric statistics [17]. The goal of adaptive inference is to construct a single procedure that achieves optimality simultaneously over a collection of parameter spaces. Informally, an adaptive procedure automatically adjusts to the *unknown* parameter, and acts as if it knows the parameter lies in a more restricted subset of the whole parameter space. A common way to evaluate such a procedure is to compare its maximum risk over each subset of the parameter space in the collection with the corresponding minimax risk. If they are nearly equal, then we say such a procedure is *adaptive* with respect to that collection of subsets of the parameter space.

The primary results of this paper are twofold.

- 1) First, we show that the minimax rate-optimal entropy estimator in Jiao et al. [11] is adaptive with respect to the collection of parameter space $\mathcal{M}_S(H)$, where $\mathcal{M}_S(H) \triangleq \{P : H(P) \leq H, P \in \mathcal{M}_S\}$. Moreover, the estimator does not need to know S nor H , which is an advantage in practice since usually the alphabet size S nor an *a priori* upper bound on the true entropy $H(P)$ are known.
- 2) Second, we show that the sample size *enlargement* effect still holds in this adaptive estimation scenario. Table I demonstrates that in estimating various functionals, the performance of the minimax rate-optimal estimator with n samples is nearly that of the MLE with $n \ln n$ samples, which the authors termed “effective sample size enlargement” in [11]. We compute the maximum risk of the MLE over each $\mathcal{M}_S(H)$, and show that for every H , the performance of the estimator in [11] with n samples is still nearly that of the MLE with $n \ln n$ samples.

These facts suggest that the estimator in Jiao et al. [11] is near *optimal* in a very strong sense, for which we refer the readers to [11] for a detailed discussion on methodology behind their estimator, literature survey, and experimental results.

C. Mathematical framework and estimator construction

Before we discuss the main results, we would like to recall the construction of the entropy estimator in [11]. The approach is to tackle the estimation problem separately for the cases of “small p ” and “large p ” in $H(P)$ estimation, corresponding to treating regions where the functional is “nonsmooth” and “smooth” in different ways. Specifically, after we obtain the empirical distribution P_n , for each coordinate $P_n(i)$, if $P_n(i) \ll \ln n/n$, we (i) compute the best polynomial approximation for $-p_i \ln p_i$ in the regime $0 \leq p_i \ll \ln n/n$, (ii) use the unbiased estimators for integer powers p_i^k to estimate the corresponding terms in the polynomial approximation for $-p_i \ln p_i$ up to order $K_n \sim \ln n$, and (iii) use that polynomial as an estimate for $-p_i \ln p_i$. If $P_n(i) \gg \ln n/n$, we use the estimator $-P_n(i) \ln P_n(i) + \frac{1}{2n}$ to estimate $-p_i \ln p_i$. Then, we add the estimators corresponding to each coordinate.

We define the minimax risk for Multinomial model with n observations on alphabet size S for estimating $H(P)$, $P \in \mathcal{M}_S(H)$ as

$$R(S, n, H) \triangleq \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_{\text{Multinomial}} \left(\hat{H} - H(P) \right)^2, \quad (3)$$

which is the quantity we will characterize in this paper. To simplify the analysis, we also utilize the Poisson sampling model, i.e., we first draw a random variable $N \sim \text{Poi}(n)$, and then obtain N samples from the distribution P . It is equivalent to having a S -dimensional random vector \mathbf{Z} such that each component Z_i in \mathbf{Z} has distribution $\text{Poi}(np_i)$, and all coordinates of \mathbf{Z} are independent.

The counterpart of minimax risk in the Poissonized model is defined as

$$R_P(S, n, H) \triangleq \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_{\text{Poisson}} \left(\hat{H} - H(P) \right)^2. \quad (4)$$

The following lemma, which follows from [11], [15], shows that the minimax risks under the Multinomial model and the Poissonized model are essentially equivalent.

Lemma 1. *The minimax risks under the Poissonized model and the Multinomial model are related via the following inequalities:*

$$R_P(S, 2n, H) - e^{-n/4} H^2 \leq R(S, n, H) \leq 2R_P(S, n/2, H). \quad (5)$$

For simplicity of analysis, we conduct the classical “splitting” operation [18] on the Poisson random vector \mathbf{Z} , and obtain two independent identically distributed random vectors $\mathbf{X} = [X_1, X_2, \dots, X_S]^T$, $\mathbf{Y} = [Y_1, Y_2, \dots, Y_S]^T$, such that each component X_i in \mathbf{X} has distribution $\text{Poi}(np_i/2)$, and all coordinates in \mathbf{X} are independent. For each coordinate i , the splitting process generates a random variable T_i such that $T_i | \mathbf{Z} \sim \text{B}(Z_i, 1/2)$, and assign $X_i = T_i$, $Y_i = Z_i - T_i$. All the random variables $\{T_i : 1 \leq i \leq S\}$ are conditionally independent given our observation \mathbf{Z} . We also note that for random variable X such that $nX \sim \text{Poi}(np)$,

$$\mathbb{E} \prod_{r=0}^{k-1} \left(X - \frac{r}{n} \right) = p^k, \quad (6)$$

for any $k \in \mathbb{N}_+$.

For simplicity, we re-define $n/2$ as n , and denote

$$\hat{p}_{i,1} = \frac{X_i}{n}, \hat{p}_{i,2} = \frac{Y_i}{n}, \Delta = \frac{c_1 \ln n}{n}, K = c_2 \ln n, t = \frac{\Delta}{4}, \quad (7)$$

where c_1, c_2 are positive parameters to be specified later. Note that Δ, K, t are functions of n , where we omit the subscript n for brevity.

The estimator \hat{H} in Jiao et al. [11] is constructed as follows.

$$\hat{H} \triangleq \sum_{i=1}^S [L_H(\hat{p}_{i,1}) \mathbb{1}(\hat{p}_{i,2} \leq 2\Delta) + U_H(\hat{p}_{i,1}) \mathbb{1}(\hat{p}_{i,2} > 2\Delta)], \quad (8)$$

where

$$S_{K,H}(x) \triangleq \sum_{k=1}^K g_{k,H}(4\Delta)^{-k+1} \prod_{r=0}^{k-1} \left(x - \frac{r}{n} \right) \quad (9)$$

$$L_H(x) \triangleq \min \{S_{K,H}(x), 1\} \quad (10)$$

$$U_H(x) \triangleq I_n(x) \left(-x \ln x + \frac{1}{2n} \right). \quad (11)$$

We explain each equation in detail as follows.

- 1) Equation (8): Note that $\hat{p}_{i,1}$ and $\hat{p}_{i,2}$ are i.i.d. random variables such that $n\hat{p}_{i,1} \sim \text{Poi}(np_i)$. We use $\hat{p}_{i,2}$ to determine whether we are operating in the “nonsmooth” regime or not. If $\hat{p}_{i,2} \leq 2\Delta$, we declare we are in the “nonsmooth” regime,

and plug in $\hat{p}_{i,1}$ into function $L_H(\cdot)$. If $\hat{p}_{i,2} > 2\Delta$, we declare we are in the “smooth” regime, and plug in $\hat{p}_{i,1}$ into $U_H(\cdot)$.

2) Equation (9):

The coefficients $r_{k,H}, 0 \leq k \leq K$ are coefficients of the best polynomial approximation of $-x \ln x$ over $[0, 1]$ up to degree K , i.e.,

$$\sum_{k=0}^K r_{k,H} x^k = \arg \min_{y(x) \in \text{poly}_K} \sup_{x \in [0,1]} |y(x) - (-x \ln x)|, \quad (12)$$

where poly_K denotes the set of algebraic polynomials up to order K . Note that in general $g_{k,\alpha}$ depends on K , which we do not make explicit for brevity.

Then we define $\{g_{k,H}\}_{1 \leq k \leq K}$

$$g_{k,H} = r_{k,H}, 2 \leq k \leq K, g_{1,H} = r_{1,H} - \ln(4\Delta). \quad (13)$$

Lemma 9 shows that for $nX \sim \text{Poi}(np)$,

$$\mathbb{E}S_{K,H}(X) = \sum_{k=1}^K g_{k,H} (4\Delta)^{-k+1} p^k \quad (14)$$

is a near-best polynomial approximation for $-p \ln p$ on $[0, 4\Delta]$. Thus, we can understand $S_{K,H}(X), nX \sim \text{Poi}(np)$ as a random variable whose expectation is nearly ¹ the best approximation of function $-x \ln x$ over $[0, 4\Delta]$.

3) Equation (10):

Any reasonable estimator for $-p \ln p$ should be upper bounded by the value one. We cutoff $S_{K,H}(x)$ by upper bound 1, and define the function $L_H(x)$, which means “lower part”.

4) Equation (11):

The function $U_H(x)$ (means “upper part”) is nothing but a product of an interpolation function $I_n(x)$ and the bias-corrected MLE. The interpolation function $I_n(x)$ is defined as follows:

$$I_n(x) = \begin{cases} 0 & x \leq t \\ g(x-t; t) & t < x < 2t \\ 1 & x \geq 2t \end{cases} \quad (15)$$

The following lemma characterizes the properties of the function $g(x; a)$ appearing in the definition of $I_n(x)$. In particular, it shows that $I_n(x) \in C^4[0, 1]$.

Lemma 2. For the function $g(x; a)$ on $[0, a]$ defined as follows,

$$g(x; a) \triangleq 126 \left(\frac{x}{a}\right)^5 - 420 \left(\frac{x}{a}\right)^6 + 540 \left(\frac{x}{a}\right)^7 - 315 \left(\frac{x}{a}\right)^8 + 70 \left(\frac{x}{a}\right)^9, \quad (16)$$

we have the following properties:

$$g(0; a) = 0, \quad g^{(i)}(0; a) = 0, 1 \leq i \leq 4 \quad (17)$$

$$g(a; a) = 1, \quad g^{(i)}(a; a) = 0, 1 \leq i \leq 4 \quad (18)$$

The function $g(x; 1)$ is depicted in Figure 1.

¹Note that we have removed the constant term from the best polynomial approximation. It is to ensure that we assign zero to symbols we do not see.

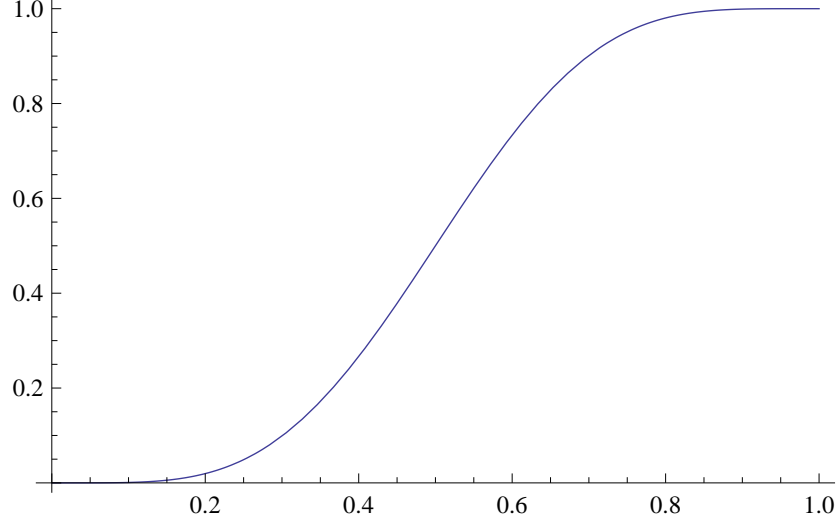


Fig. 1: The function $g(x; 1)$ over interval $[0, 1]$.

II. MAIN RESULTS

Since $\sup_{P \in \mathcal{M}_S} H(P) = \ln S$, we assume throughout this paper that $0 < H \leq \ln S$. Denote by $\mathcal{M}_S(H)$ the set of all discrete probability distributions P with support size $|\text{supp}(P)| = S$ and entropy $H(P) \leq H$. We say an estimator $\hat{H} \equiv \hat{H}(\mathbf{Z})$ is within accuracy $\epsilon > 0$, if and only if

$$\sup_{P \in \mathcal{M}_S(H)} \left(\mathbb{E}_P |\hat{H} - H(P)|^2 \right)^{\frac{1}{2}} \leq \epsilon. \quad (19)$$

For the plug-in estimator $H(P_n)$, the following theorem presents the non-asymptotic upper and lower bounds for the L_2 risk.

Theorem 1. *If $H \geq H_0 > 0$, where H_0 is a universal positive constant, then for the plug-in estimator $H(P_n)$, we have*

$$\sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P |H(P_n) - H(P)|^2 \asymp \begin{cases} \left(\frac{S}{n}\right)^2 + \frac{H \ln S}{n} & \text{if } S \ln S \leq enH, \\ \left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH}\right)\right]^2 & \text{otherwise.} \end{cases} \quad (20)$$

Note that the only assumption in Theorem 1 is that the upper bound H should be no smaller than a constant, which is a reasonable assumption to avoid the subtle case where the naive zero estimator $\hat{H} \equiv 0$ has a satisfactory performance. The minimum sample complexity of the plug-in approach can be immediately obtained from Theorem 1.

Corollary 1. *If $H \geq H_0 > 0$, where H_0 is a universal positive constant, the plug-in estimator $H(P_n)$ is within accuracy ϵ if and only if $n \succeq (S^{1-\frac{\epsilon}{H}} \cdot \frac{\ln S}{H})$.*

Recall that it requires $n \succeq (\frac{S}{\epsilon})$ samples for the MLE to achieve accuracy ϵ when there is no constraint on the entropy [11]. Hence, when the upper bound on the entropy is loose, i.e., $H \asymp \ln S$, the minimum sample complexity in the bounded entropy case is exactly the same, i.e., we cannot essentially improve the estimation performance. On the other hand, when the upper bound is tight, i.e., $H \ll \ln S$, the required sample complexity enjoyed a significant reduction, i.e., we only need a sublinear number of samples for accurate entropy estimation.

When it comes to the maximum L_2 risk, we conclude from Theorem 1 that the bounded entropy property helps only at the boundary, i.e., when n is close to S and H is small. Moreover, this help vanishes quickly as S increases: when $n = S^{1-\delta}$, the maximum L_2 risk will be at the order $(\delta H)^2$, which is the same risk achieved by the naive zero estimator when δ is not close to zero.

Is the plug-in estimator $H(P_n)$ optimal in the minimax sense? It has been shown in [11], [12], [15] that when there is no constraint on $H(P)$, i.e., $H = \ln S$, the answer is *negative*. What about subsets of \mathcal{M}_S , such as $\mathcal{M}_S(H)$? The following theorem characterizes the minimax L_2 rates over $\mathcal{M}_S(H)$.

Theorem 2. *If $H \geq H_0 > 0$, where H_0 is a universal positive constant, then*

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P |\hat{H} - H(P)|^2 \asymp \begin{cases} \frac{S^2}{(n \ln n)^2} + \frac{H \ln S}{n} & \text{if } S \ln S \leq enH \ln n, \\ \left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n}\right)\right]^2 & \text{otherwise.} \end{cases} \quad (21)$$

where the infimum is taken over all possible estimators. Moreover, the upper bound is achieved by the estimator in [11] under the Poissonized model without the knowledge of H nor S .

An immediate result on the sample complexity is as follows.

Corollary 2. *If $H \geq H_0 > 0$, where H_0 is a universal positive constant, the minimax rate-optimal estimator in [11] is within accuracy ϵ if and only if $n \succeq \left(\frac{1}{H} S^{1-\frac{\epsilon}{H}}\right)$.*

For the minimum sample complexity, we still distinguish H into two cases. Firstly, when $H \asymp \ln S$, the required sample complexity is $n \asymp \frac{S}{\epsilon \ln S}$, which recovers the minimax results with no constraint on entropy in [11]. Secondly, when $H \ll \ln S$, there is a significant improvement.

We also conclude from Theorem 2 that the bounded entropy constraint again helps only at the boundary, and this help vanishes quickly as S increases: when $n = S^{1-\delta}$, we do not have sufficient information to make inference, and the naive zero estimator is near-minimax.

To sum up, we have obtained the following conclusions.

- 1) The minimax rate-optimal entropy estimator in Jiao et al. [11] is adaptive with respect to the collection of parameter space $\mathcal{M}_S(H)$, where $\mathcal{M}_S(H) \triangleq \{P : H(P) \leq H, P \in \mathcal{M}_S\}$. Moreover, the estimator does not need to know S nor H , which is an advantage in practice since usually the alphabet size S nor an *a priori* upper bound on the true entropy $H(P)$ are known.
- 2) Second, the sample size *enlargement* effect still holds in this adaptive estimation scenario. Table I demonstrates that in estimating various functionals, the performance of the minimax rate-optimal estimator with n samples is essentially that of the MLE with $n \ln n$ samples, which the authors termed ‘‘sample size enlargement’’ in [11]. Theorems 1 and 2 show that over every $\mathcal{M}_S(H)$, the performance of the estimator in [11] with n samples is still essentially that of the MLE with $n \ln n$ samples.

III. PROOF OF UPPER BOUNDS IN THEOREM 1

First we consider the case where $S \ln S \leq enH$. For the bias, it has been shown in [13] that

$$\text{Bias}(H(P_n)) \leq \ln \left(1 + \frac{S-1}{n}\right) \leq \frac{S}{n}. \quad (22)$$

As for the variance, [11] shows that by the Efron-Stein inequality that

$$\text{Var}(H(P_n)) \leq \frac{2}{n} \sum_{i=1}^S p_i (\ln p_i - 2)^2 \leq \frac{2}{n} \left(\sum_{i=1}^S p_i (\ln p_i)^2 + 4H + 4 \right). \quad (23)$$

Lemma 3. *For any discrete distribution $P = (p_1, p_2, \dots, p_S)$ with alphabet size $S \geq 2$, we have*

$$\sum_{i=1}^S p_i (\ln p_i)^2 \leq 2 \ln S \cdot \left(\sum_{i=1}^S -p_i \ln p_i \right) + 3. \quad (24)$$

In light of Lemma 3, we conclude that

$$\text{Var}(H(P_n)) \leq \frac{2}{n} \left(\sum_{i=1}^S p_i (\ln p_i)^2 + 4H + 4 \right) \leq \frac{2}{n} (2H \ln S + 4H + 7) \preceq \frac{H \ln S}{n} \quad (25)$$

where we have used the assumption $H \geq H_0 > 0$ in the last step.

Hence, when $S \ln S \leq enH$, we have

$$\mathbb{E}_P (H(P_n) - H(P))^2 = (\text{Bias}(H(P_n)))^2 + \text{Var}(H(P_n)) \preceq \frac{S^2}{n^2} + \frac{H \ln S}{n} \quad (26)$$

which completes the proof for the first part. For the second part, we introduce a lemma first.

Lemma 4. *For $p \leq \frac{c}{2en}$ and $n\hat{p} \sim \mathcal{B}(n, p)$, where c is a positive integer, we have*

$$0 \leq -p \ln p - \mathbb{E}[-\hat{p} \ln \hat{p}] \leq -p \ln(np) + p \ln c + \frac{c \ln c}{n} \left(\frac{enp}{c}\right)^c + \sum_{k=c+1}^n \frac{\ln k + 1}{n} \left(\frac{enp}{k}\right)^k. \quad (27)$$

Define $\xi(\hat{p}_i) = -\hat{p}_i \ln \hat{p}_i$. In light of Lemma 4, for any positive integer c , we have

$$\sum_{i: p_i \leq \frac{c}{2en}} |\text{Bias}(\xi(\hat{p}_i))| \leq \sum_{i: p_i \leq \frac{c}{2en}} \left[-p_i \ln(np_i) + p_i \ln c + \frac{c \ln c}{n} \left(\frac{enp_i}{c} \right)^c + \sum_{k=c+1}^n \frac{\ln k + 1}{n} \left(\frac{enp_i}{k} \right)^k \right] \quad (28)$$

$$\leq \sum_{i: p_i \leq \frac{c}{2en}} \left[-p_i \ln \left(\frac{np_i}{c} \right) \right] + \sum_{i: p_i \leq \frac{c}{2en}} \left[\frac{c \ln c}{n} \left(\frac{enp_i}{c} \right)^c + \sum_{k=c+1}^n \frac{\ln k + 1}{n} \left(\frac{enp_i}{k} \right)^k \right] \quad (29)$$

$$\leq \sum_{i: p_i \leq \frac{c}{2en}} \left[-p_i \ln \left(\frac{np_i}{c} \right) \right] + \frac{c \ln c}{n} \cdot \frac{2en}{c} 2^{-c} + \sum_{k=c+1}^n \frac{\ln k + 1}{n} \cdot \frac{2en}{c} \left(\frac{c}{2k} \right)^k \quad (30)$$

$$\leq \sum_{i: p_i \leq \frac{c}{2en}} \left[-p_i \ln \left(\frac{np_i}{c} \right) \right] + 2e \cdot 2^{-c} \ln c + \sum_{k=c+1}^n 2e \cdot \frac{\ln k + 1}{k 2^k} \quad (31)$$

$$\leq \sum_{i: p_i \leq \frac{c}{2en}} \left[-p_i \ln \left(\frac{np_i}{c} \right) \right] + 2e \cdot 2^{-c} (\ln c + 1), \quad (32)$$

where we have used the convexity of x^k , $0 \leq x \leq 1$ for any $k \geq 1$ in (30). We consider the following optimization problem:

$$\text{maximize } \sum_{i: p_i \leq \frac{c}{2en}} -p_i \ln \left(\frac{np_i}{c} \right) \quad \text{subject to } \sum_{i: p_i \leq \frac{c}{2en}} -p_i \ln p_i \leq H, A_1 \equiv \left| \left\{ i : p_i \leq \frac{c}{2en} \right\} \right| \leq S \quad (33)$$

It is straightforward to show that in the solution to (33), all $p_i \leq c/2en$ should be equal, say, to p_0 . Then (33) reduces to

$$\text{maximize } A_1 p_0 \ln \left(\frac{c}{np_0} \right) \quad \text{subject to } 0 \leq p_0 \leq \frac{c}{2en}, A_1 \leq S, -A_1 p_0 \ln p_0 \leq H \quad (34)$$

whose optimization result is no larger than

$$\text{maximize } A_1 p_0 \ln \left(\frac{c}{np_0} \right) \quad \text{subject to } A_1 \leq S, -A_1 p_0 \ln p_0 \leq H. \quad (35)$$

Then it is easy to check that the solution to (35) is $A_1 = S$ and $-A_1 p_0 \ln p_0 = H$. Then we have $p_0 \asymp \frac{H}{S \ln S}$, and

$$\sum_{i: p_i \leq \frac{c}{2en}} |\text{Bias}(\xi(\hat{p}_i))| \leq S p_0 \ln \left(\frac{c}{np_0} \right) + 2e \cdot 2^{-c} (\ln c + 1) \quad (36)$$

$$\preceq \frac{H}{\ln S} \ln \left(\frac{c S \ln S}{n H} \right) + 2^{-c} \ln c. \quad (37)$$

Now we set $c = n^{\frac{\epsilon}{H}}$ with

$$\epsilon = \frac{H}{\ln S} \ln \left(\frac{S \ln S}{n H} \right) > 0 \quad (38)$$

and we assume without loss of generality that c is an integer. We can easily check that

$$2^{-c} \ln c \preceq \frac{H \ln c}{\ln S} = \frac{\ln n}{\ln S} \epsilon \preceq \epsilon = \frac{H}{\ln S} \ln \left(\frac{S \ln S}{n H} \right) \quad (39)$$

which leads to the desired result

$$\sum_{i: p_i \leq \frac{c}{2en}} |\text{Bias}(\xi(\hat{p}_i))| \preceq \frac{H}{\ln S} \ln \left(\frac{S \ln S}{n H} \right). \quad (40)$$

As for the second part of bias, it has been shown in [11] that $|\text{Bias}(\xi(\hat{p}_i))| \leq \frac{5 \ln 2}{n}$ holds for all i , hence

$$\sum_{i: p_i > \frac{c}{2en}} |\text{Bias}(\xi(\hat{p}_i))| \leq \frac{5 \ln 2}{n} \cdot \left| \left\{ i : p_i > \frac{c}{2en} \right\} \right| \equiv \frac{5 \ln 2}{n} \cdot A_2. \quad (41)$$

We use the bounded entropy property to bound A_2 . Due to the concavity of $-x \ln x$, $0 \leq x \leq 1$, the minimum of $\sum_{i: p_i > \frac{c}{2en}} -p_i \ln p_i$ is attained when all but one p_i are at the boundary $p_i = \frac{c}{2en}$, hence

$$H \geq \sum_{i=1}^S -p_i \ln p_i \geq \sum_{i: p_i > \frac{c}{2en}} -p_i \ln p_i \geq (A_2 - 1) \cdot \frac{c}{2en} \ln \left(\frac{2en}{c} \right). \quad (42)$$

As a result, we have $A_2 \preceq \frac{nH}{\ln n}$, and

$$\sum_{i: p_i > \frac{c}{2en}} |\text{Bias}(\xi(\hat{p}_i))| \preceq \frac{H}{\ln n} \preceq \frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH} \right) \quad (43)$$

when $S \ln S \geq enH$ (the last inequality can be shown by considering two cases $S \geq n^2$ and $S < n^2$ separately). Hence,

$$|\text{Bias}(H(P_n))| \leq \sum_{i: p_i \leq \frac{c}{2en}} |\text{Bias}(\xi(\hat{p}_i))| + \sum_{i: p_i > \frac{c}{2en}} |\text{Bias}(\xi(\hat{p}_i))| \preceq \frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH} \right) \quad (44)$$

and the squared bias is the dominating term since

$$\left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH} \right) \right]^2 \succeq \frac{(\ln n)^2}{n} \quad (45)$$

where $\frac{(\ln n)^2}{n}$ is an upper bound for the variance $\text{Var}(H(P_n))$ [11].

IV. PROOF OF LOWER BOUNDS IN THEOREM 1

We first derive a lower bound for the bias term. When $S \ln S \leq enH$, we recall the following result in [11].

Lemma 5. For $p \geq \frac{15}{n}, p \in [0, 1]$, we have

$$-p \ln p - \mathbb{E}[-\hat{p} \ln \hat{p}] \geq \frac{1-p}{2n} + \frac{1}{20n^2 p} - \frac{p}{12n^2}. \quad (46)$$

If we choose $P = (\frac{15}{n}, \frac{15}{n}, \dots, \frac{15}{n}, 1 - \frac{15K}{n}, 0, \dots, 0)$, where

$$K = \min \left\{ \left\lfloor \frac{n}{15} \right\rfloor, S-1, \max \left\{ N \in \mathbb{N} : -\frac{15N}{n} \ln \left(\frac{15}{n} \right) - \left(1 - \frac{15N}{n} \right) \ln \left(1 - \frac{15N}{n} \right) \leq H \right\} \right\} \quad (47)$$

we have

$$|\text{Bias}(H(P_n))| \geq K \cdot \left(\frac{n-15}{2n^2} + \frac{1}{300n} - \frac{5}{4n^3} \right) \succeq \min \left\{ 1, \frac{S}{n}, \frac{H}{\ln n} \right\} \succeq \frac{S}{n} \quad (48)$$

where we have used the assumption $S \ln S \leq enH$ and $H \geq H_0 > 0$. Hence, we have proved that

$$\sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P |H(P_n) - H(P)|^2 \succeq \frac{S^2}{n^2}. \quad (49)$$

For the case where $S \ln S > enH$, we establish a lemma first.

Lemma 6. For $n\hat{p} \sim \text{B}(n, p)$, we have

$$-p \ln p - \mathbb{E}[-\hat{p} \ln \hat{p}] \geq -p \ln(np). \quad (50)$$

Proof: Note that $n\hat{p}$ is an integer, we have

$$\mathbb{E}[-\hat{p} \ln \hat{p}] \leq \mathbb{E}[\hat{p} \ln n] = p \ln n. \quad (51)$$

Consider a distribution $P = (\frac{A}{S-1}, \dots, \frac{A}{S-1}, A) \in \mathcal{M}_S$, where $A \in (0, 1)$ is the solution to

$$-A \ln \left(\frac{A}{S-1} \right) - (1-A) \ln(1-A) = H \quad (52)$$

then Lemma 6 tells us that

$$|\text{Bias}(H(P_n))| \geq \sum_{i=1}^{S-1} -p_i \ln(np_i) = -A \ln \left(\frac{nA}{S-1} \right) \succeq \frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH} \right) \quad (53)$$

where in the last step we have used the relationship $A \asymp \frac{H}{\ln S}$ from (52).

We now turn to the lower bound for variance. We will actually prove a stronger result: a minimax lower bound for all estimators for the L_2 risk, which naturally is also a lower bound for the maximum risk of the MLE. We use Le Cam's two-point method here. Suppose we observe a random vector $\mathbf{Z} \in (\mathcal{Z}, \mathcal{A})$ which has distribution P_θ where $\theta \in \Theta$. Let θ_0 and θ_1 be two elements of Θ . Let $\hat{T} = \hat{T}(\mathbf{Z})$ be an arbitrary estimator of a function $T(\theta)$ based on \mathbf{Z} . We have the following general minimax lower bound.

Lemma 7. [19, Sec. 2.4.2] Denoting the Kullback-Leibler divergence between P and Q by

$$D(P\|Q) \triangleq \begin{cases} \int \ln\left(\frac{dP}{dQ}\right) dP, & \text{if } P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases} \quad (54)$$

we have

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbb{P}_{\theta} \left(|\hat{T} - T(\theta)| \geq \frac{|T(\theta_1) - T(\theta_0)|}{2} \right) \geq \frac{1}{4} \exp(-D(P_{\theta_1}\|P_{\theta_0})). \quad (55)$$

Applying this lemma to the Poissonized model $n\hat{p}_i \sim \text{Poi}(np_i)$, $1 \leq i \leq S$, we know that for $\theta_1 = (p_1, p_2, \dots, p_S)$, $\theta_0 = (q_1, q_2, \dots, q_S)$,

$$D(P_{\theta_1}\|P_{\theta_0}) = \sum_{i=1}^S D(\text{Poi}(np_i)\|\text{Poi}(nq_i)) = \sum_{i=1}^S \sum_{k=0}^{\infty} \mathbb{P}(\text{Poi}(np_i) = k) \cdot k \ln \frac{p_i}{q_i} = \sum_{i=1}^S np_i \ln \frac{p_i}{q_i} = nD(\theta_1\|\theta_0), \quad (56)$$

then Markov's inequality yields

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2 \geq \frac{|H(\theta_1) - H(\theta_0)|^2}{4} \cdot \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{P} \left(|\hat{H} - H(P)| \geq \frac{|H(\theta_1) - H(\theta_0)|}{2} \right) \quad (57)$$

$$\geq \frac{|H(\theta_1) - H(\theta_0)|^2}{16} \exp(-nD(\theta_1\|\theta_0)). \quad (58)$$

Fix $\epsilon \in (0, 1)$ to be specified later, and let

$$\theta_1 = \left(\frac{A}{S-1}, \dots, \frac{A}{S-1}, 1-A \right), \quad \theta_0 = \left(\frac{A(1-\epsilon)}{S-1}, \dots, \frac{A(1-\epsilon)}{S-1}, 1-A+A\epsilon \right), \quad (59)$$

where A is the solution to (52). Direct computation yields

$$D(\theta_1\|\theta_0) = A \ln \frac{1}{1-\epsilon} + (1-A) \ln \frac{1-A}{1-A+A\epsilon} \equiv h(\epsilon), \quad (60)$$

we have $h(0) = h'(0) = 0$, and $|h''(0)| = \frac{1-A}{A} > 0$. Hence, for ϵ small enough we have $D(\theta_1\|\theta_0) \leq \epsilon^2/A$. By choosing $\epsilon = (nA)^{-\frac{1}{2}} \leq 1$, we have

$$|H(\theta_1) - H(\theta_0)| = \left| -A \ln \left(\frac{A}{S-1} \right) + A(1-\epsilon) \ln \left(\frac{A(1-\epsilon)}{S-1} \right) - (1-A) \ln(1-A) + (1-A+A\epsilon) \ln(1-A+A\epsilon) \right| \quad (61)$$

$$\geq A\epsilon \ln \left(\frac{S-1}{A} \right). \quad (62)$$

Hence, by Lemma 7 we know that

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P |\hat{H} - H(P)|^2 \geq \left[A\epsilon \ln \left(\frac{S-1}{A} \right) \right]^2 \asymp \frac{H}{n \ln S} \left[\ln \left(\frac{S \ln S}{H} \right) \right]^2 \asymp \frac{H \ln S}{n} \quad (63)$$

and the lower bound in the Multinomial model follows from Lemma 1.

V. PROOF OF UPPER BOUNDS IN THEOREM 2

Define

$$\xi \triangleq \xi(X, Y) = L_H(X) \mathbb{1}(Y \leq 2\Delta) + U_H(X) \mathbb{1}(Y > 2\Delta), \quad (64)$$

where $nX \stackrel{D}{=} nY \sim \text{Poi}(np)$, and X, Y are independent. We first recall the following lemma from [11].

Lemma 8. Suppose $0 < c_1 = 16(1 + \delta)$, $0 < 8c_2 \ln 2 = \epsilon < 1$, $\delta > 0$. Then the bias and variance of $\xi(X, Y)$ are given as follows:

$$|\text{Bias}(\xi)| \leq \frac{1}{n \ln n} \quad (65)$$

$$\text{Var}(\xi) \leq \frac{(\ln n)^4}{n^{2-\epsilon}} + \frac{p(\ln p)^2}{n} \quad (66)$$

In light of Lemma 10, we have

$$|\text{Bias}(\hat{H})| \leq \sum_{i=1}^S |\text{Bias}(\xi(\hat{p}_{i,1}, \hat{p}_{i,2}))| \leq \sum_{i=1}^S \frac{1}{n \ln n} = \frac{S}{n \ln n} \quad (67)$$

$$\text{Var}(\hat{H}) = \sum_{i=1}^S \text{Var}(\xi(\hat{p}_{i,1}, \hat{p}_{i,2})) \leq \sum_{i=1}^S \left(\frac{(\ln n)^4}{n^{2-\epsilon}} + \frac{p_i (\ln p_i)^2}{n} \right) \leq \frac{S(\ln n)^4}{n^{2-\epsilon}} + \frac{H \ln S}{n} \quad (68)$$

where we have used Lemma 3 in the last step. Hence,

$$\mathbb{E}_P \left(\hat{H} - H(P) \right)^2 = |\text{Bias}(\hat{H})|^2 + \text{Var}(\hat{H}) \leq \frac{S^2}{(n \ln n)^2} + \frac{S(\ln n)^4}{n^{2-\epsilon}} + \frac{H \ln S}{n}. \quad (69)$$

When $S \ln S \leq enH \ln n$, for ϵ small enough, say, $\epsilon < \frac{1}{2}$, we have

$$\frac{S(\ln n)^4}{n^{2-\epsilon}} \leq \sqrt{\frac{S^2}{(n \ln n)^2} \cdot \frac{H \ln S}{n}} \leq \frac{S^2}{(n \ln n)^2} + \frac{H \ln S}{n} \quad (70)$$

where we have used the assumption that $H \geq H_0 > 0$. Hence, the term $\frac{S(\ln n)^4}{n^{2-\epsilon}}$ is negligible when compared with others, and we have reached the end for the case $S \ln S \leq enH \ln n$.

For the case where $S \ln S \geq enH \ln n$, we need stronger results for the bias and variance in the regime where $p < \frac{1}{en \ln n}$. The results are summarized in the following lemma.

Lemma 9. *If $0 < c_2 \leq 1 \leq c_1$, for $nX \sim \text{Poi}(np)$, $0 < p < \frac{1}{en \ln n}$, we have*

$$|\mathbb{E} S_{K,H}(X) + p \ln p| \leq -p \ln(pn \ln n) + (D_p + \ln(4c_1/c_2^2))p \quad (71)$$

$$\mathbb{E} S_{K,H}^2(X) \leq 2^{10c_2 \ln 2} \frac{(4c_1 \ln n)^4 p}{n} \quad (72)$$

where the constant D_p is given in Lemma 15.

Using the Poisson tail bound (cf. Lemma 17) and similar argument to [11, Lem. 8], we have the following lemma.

Lemma 10. *Suppose $0 < c_1 = 16(1 + \delta)$, $0 < 10c_2 \ln 2 = \epsilon < 1$, $\delta > 0$. Then for $0 < p < \frac{1}{en \ln n}$, we have*

$$|\text{Bias}(\xi)| \leq -p \ln(pn \ln n) \quad (73)$$

$$\text{Var}(\xi) \leq \frac{(\ln n)^4 p}{n^{1-\epsilon}} \quad (74)$$

Now we proceed to bound the total bias and variance. By looking at the maximization problem

$$\sum_{i: p_i < \frac{1}{en \ln n}} -p_i \ln(p_i n \ln n) \quad \text{subject to} \quad \sum_{i: p_i < \frac{1}{en \ln n}} -p_i \ln p_i \leq H, \left| \left\{ i : p_i < \frac{1}{en \ln n} \right\} \right| \leq S. \quad (75)$$

Using similar arguments to (33), all $p_i \leq \frac{1}{en \ln n}$ should be equal and both equalities in the constraints hold. As a result, we have

$$\sum_{i: p_i < \frac{1}{en \ln n}} |\text{Bias}(\xi)| \leq \sum_{i: p_i < \frac{1}{en \ln n}} -p_i \ln(p_i n \ln n) \leq \frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n} \right). \quad (76)$$

For symbols with $p_i > \frac{1}{en \ln n}$, similar arguments in the MLE analysis yield

$$\left| \left\{ i : p_i \geq \frac{1}{en \ln n} \right\} \right| \leq \frac{nH \ln n}{\ln(n \ln n)} \asymp nH \quad (77)$$

hence

$$\sum_{i: p_i \geq \frac{1}{en \ln n}} |\text{Bias}(\xi)| \leq \frac{1}{n \ln n} \cdot \left| \left\{ i : p_i \geq \frac{1}{en \ln n} \right\} \right| \leq \frac{H}{\ln n} \leq \frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n} \right) \quad (78)$$

when $S \ln S \geq enH \ln n$. Summing up the bias yields

$$|\text{Bias}(\hat{H})| \leq \sum_{i: p_i \geq \frac{1}{en \ln n}} |\text{Bias}(\xi(\hat{p}_{i,1}, \hat{p}_{i,2}))| + \sum_{i: p_i < \frac{1}{en \ln n}} |\text{Bias}(\xi(\hat{p}_{i,1}, \hat{p}_{i,2}))| \leq \frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n} \right). \quad (79)$$

For the total variance, we have

$$\text{Var}(\hat{H}) = \sum_{i: p_i \geq \frac{1}{en \ln n}} \text{Var}(\xi(\hat{p}_{i,1}, \hat{p}_{i,2})) + \sum_{i: p_i < \frac{1}{en \ln n}} \text{Var}(\xi(\hat{p}_{i,1}, \hat{p}_{i,2})) \quad (80)$$

$$\preceq \sum_{i: p_i \geq \frac{1}{en \ln n}} \left(\frac{(\ln n)^4}{n^{2-\epsilon}} + \frac{p(\ln p)^2}{n} \right) + \sum_{i: p_i < \frac{1}{en \ln n}} \frac{(\ln n)^4 p}{n^{1-\epsilon}} \quad (81)$$

$$\preceq \left(\frac{(\ln n)^5}{n^{1-\epsilon}} + \frac{(\ln n)^2}{n} \right) + \frac{(\ln n)^4}{n^{1-\epsilon}} \quad (82)$$

$$\preceq \left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{n H \ln n} \right) \right]^2 \quad (83)$$

where in the last step we have used the assumption $H \geq H_0 > 0$ again. Combining the total bias and variance constitutes a complete proof of the upper bounds in Theorem 2.

VI. PROOF OF LOWER BOUNDS IN THEOREM 2

When $S \ln S \leq enH \ln n$, the lower bound for the squared bias, i.e., the $\frac{S^2}{(n \ln n)^2}$ term, can be obtained using a similar argument in [15]. Specifically, we can assign two product measures μ_0^N and μ_1^N to the first $N (\leq S)$ components in the distribution vector P , where

$$\text{supp}(\mu_i) = \{0\} \cup \left[\frac{1}{a_1 n \ln n}, \frac{a_2 \ln n}{n} \right], \quad i = 0, 1 \quad (84)$$

for some constants $a_1, a_2 > 0$, and

$$\int_0^1 t \mu_i(dt) = \frac{1}{a_1 n \ln n}, \quad i = 0, 1. \quad (85)$$

In particular,

$$\int_0^1 -t \ln t \mu_1(dt) - \int_0^1 -t \ln t \mu_0(dt) \succeq \frac{1}{n \ln n} \quad (86)$$

and

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2 \succeq \left[N \left(\int_0^1 -t \ln t \mu_1(dt) - \int_0^1 -t \ln t \mu_0(dt) \right) \right]^2 \succeq \frac{N^2}{(n \ln n)^2}. \quad (87)$$

In [15], $N = S$. However, in our case, we have an additional constraint that $H(P) \leq H$. Since

$$\mathbb{E}_{\mu_i}[-p \ln p] = \int_0^1 -t \ln t \mu_i(dt) \leq \ln(a_1 n \ln n) \int_0^1 t \mu_i(dt) = \frac{a_1 \ln(a_1 n \ln n)}{n \ln n} \asymp \frac{1}{n} \quad (88)$$

we have

$$\mathbb{E}_{\mu_i^N} H(P) = N \mathbb{E}_{\mu_i}[-p \ln p] \asymp \frac{N}{n}. \quad (89)$$

One can show that the measures $\mu_i^N, i = 0, 1$ are highly concentrated around their expectations [15]. Hence, in order to ensure $H(P) \leq H$ with overwhelming probability, we can set $N \asymp \min\{nH, S\}$, and the condition $S \ln S \leq enH \ln n$ and $H \geq H_0 > 0$ yield that $N \succeq S$. Hence,

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2 \succeq \frac{N^2}{(n \ln n)^2} \succeq \frac{S^2}{(n \ln n)^2}. \quad (90)$$

The variance bound $\frac{H \ln S}{n}$ has been given in (63), and so far we have completed the proof of the first part. As for the second part, the key lemma we will employ is the so-called method of two fuzzy hypotheses presented in Tsybakov [19]. Below we briefly review this general minimax lower bound.

Suppose we observe a random vector $\mathbf{Z} \in (\mathcal{Z}, \mathcal{A})$ which has distribution P_θ where $\theta \in \Theta$. Let σ_0 and σ_1 be two prior distributions supported on Θ . Write F_i for the marginal distribution of \mathbf{Z} when the prior is σ_i for $i = 0, 1$. For any function g we shall write $\mathbb{E}_{F_i} g(\mathbf{Z})$ for the expectation of $g(\mathbf{Z})$ with respect to the marginal distribution of \mathbf{Z} when the prior on θ is σ_i . We shall write $\mathbb{E}_\theta g(\mathbf{Z})$ for the expectation of $g(\mathbf{Z})$ under P_θ . Let $\hat{T} = \hat{T}(\mathbf{Z})$ be an arbitrary estimator of a function $T(\theta)$ based on \mathbf{Z} . We have the following general minimax lower bound.

Lemma 11. [19, Thm. 2.15] Given the setting above, suppose there exist $\zeta \in \mathbb{R}, s > 0, 0 \leq \beta_0, \beta_1 < 1$ such that

$$\sigma_0(\theta : T(\theta) \leq \zeta - s) \geq 1 - \beta_0 \quad (91)$$

$$\sigma_1(\theta : T(\theta) \geq \zeta + s) \geq 1 - \beta_1. \quad (92)$$

If $V(F_1, F_0) \leq \eta < 1$, then

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(|\hat{T} - T(\theta)| \geq s \right) \geq \frac{1 - \eta - \beta_0 - \beta_1}{2}, \quad (93)$$

where $F_i, i = 0, 1$ are the marginal distributions of \mathbf{Z} when the priors are $\sigma_i, i = 0, 1$, respectively.

Here $V(P, Q)$ is the total variation distance between two probability measures P, Q on the measurable space $(\mathcal{Z}, \mathcal{A})$. Concretely, we have

$$V(P, Q) \triangleq \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \frac{1}{2} \int |p - q| d\nu, \quad (94)$$

where $p = \frac{dP}{d\nu}, q = \frac{dQ}{d\nu}$, and ν is a dominating measure so that $P \ll \nu, Q \ll \nu$.

First we assume that $S \leq n^{\frac{3}{2}}$. In light of Lemma 11, we construct two measures as follows.

Lemma 12. For any $0 < \eta < 1$ and positive integer $L > 0$, there exist two probability measures ν_0 and ν_1 on $[\eta, 1]$ such that

- 1) $\int t^l \nu_1(dt) = \int t^l \nu_0(dt)$, for all $l = 0, 1, 2, \dots, L$;
- 2) $\int -\ln t \nu_1(dt) - \int -\ln t \nu_0(dt) = 2E_L[-\ln x]_{[\eta, 1]}$,

where $E_L[-\ln x]_{[\eta, 1]}$ is the distance in the uniform norm on $[\eta, 1]$ from the function $f(x) = -\ln x$ to the space spanned by $\{1, x, \dots, x^L\}$.

Based on Lemma 12, two new measures $\tilde{\nu}_0, \tilde{\nu}_1$ can be constructed as follows: for $i = 0, 1$, the restriction of $\tilde{\nu}_i$ on $[\eta, 1]$ is absolutely continuous with respect to ν_i , with the Radon-Nikodym derivative given by

$$\frac{d\tilde{\nu}_i}{d\nu_i}(t) = \frac{\eta}{t}, \quad t \in [\eta, 1], \quad (95)$$

and $\tilde{\nu}_i(\{0\}) = 1 - \tilde{\nu}_i([\eta, 1]) \geq 0$. Hence, $\tilde{\nu}_0, \tilde{\nu}_1$ are both probability measures on $[0, 1]$, with the following properties

- 1) $\int t^1 \tilde{\nu}_1(dt) = \int t^1 \tilde{\nu}_0(dt) = \eta$;
- 2) $\int t^l \tilde{\nu}_1(dt) = \int t^l \tilde{\nu}_0(dt)$, for all $l = 2, \dots, L + 1$;
- 3) $\int -t \ln t \tilde{\nu}_1(dt) - \int -t \ln t \tilde{\nu}_0(dt) = 2\eta E_L[-\ln x]_{[\eta, 1]}$.

The construction of measures $\tilde{\nu}_0, \tilde{\nu}_1$ are inspired by Wu and Yang [15].

The following lemma characterizes the properties of $E_L[-\ln x]_{[\eta, 1]}$.

Lemma 13. If $K \geq eL^2$, there exists a universal constant $D_0 \geq 1$ such that

$$E_L[-\ln x]_{[(D_0 K)^{-1}, 1]} \geq \ln \left(\frac{K}{L^2} \right). \quad (96)$$

Define

$$L = d_2 \ln n, \quad \eta = \frac{nH}{d_2^2 D_0 S \ln S \ln n}, \quad M = \frac{H}{\ln S} \cdot \frac{d_1}{S\eta} = \frac{d_1 d_2^2 D_0 \ln n}{n}, \quad (97)$$

with universal positive constants $d_1 \in (0, e^{-1}]$, $d_2 > 2$ to be determined later. Without loss of generality we assume that $d_2 \ln n$ is always a positive integer. Due to $S \ln S \geq enH \ln n$, we have $(D_0 \eta)^{-1} \geq eL^2$, thus Lemma 13 yields

$$E_L[-\ln x]_{[\eta, 1]} \geq \ln \left(\frac{1}{D_0 \eta L^2} \right) \geq \ln \left(\frac{S \ln S}{H n \ln n} \right). \quad (98)$$

Let $g(x) = Mx$ and let μ_i be the measures on $[0, M]$ defined by $\mu_i(A) = \tilde{\nu}_i(g^{-1}(A))$ for $i = 0, 1$. It then follows that

- 1) $\int t^1 \mu_1(dt) = \int t^1 \mu_0(dt) = d_1 H / (S \ln S)$;
- 2) $\int t^l \mu_1(dt) = \int t^l \mu_0(dt)$, for all $l = 2, \dots, L + 1$;
- 3) $\int -t \ln t \mu_1(dt) - \int -t \ln t \mu_0(dt) = 2\eta M E_L[-\ln x]_{[\eta, 1]}$.

Let μ_0^{S-1} and μ_1^{S-1} be product priors which we assign to the length- $(S-1)$ vector $(p_1, p_2, \dots, p_{S-1})$, and we set $p_S = d_1(1 - H/\ln S)$. With a little abuse of notation, we still denote the overall product measure by μ_0^S and μ_1^S . Note that P may not be a probability distribution, we consider the set of *approximate* probability vectors

$$\mathcal{M}_S(\epsilon, H) \triangleq \left\{ P : \left| \sum_{i=1}^S p_i - d_1 \right| \leq \epsilon, H(P) \leq H, p_i \geq 0 (1 \leq i \leq S) \right\}, \quad (99)$$

with parameter $\epsilon > 0$ to be specified later, and further define under the Poissonized model,

$$R_P(S, n, H, \epsilon) \triangleq \inf_{\hat{F}} \sup_{P \in \mathcal{M}_S(\epsilon, H)} \mathbb{E}_P |\hat{H} - H(P)|^2. \quad (100)$$

Lemma 14. *For any $S, n \in \mathbb{N}$ and $0 < \epsilon < d_1$, we have*

$$R(S, n, H) \geq \frac{1}{2d_1^2} R_P \left(S, \frac{2n}{d_1}, \ln(d_1 - \epsilon) (H - \ln(d_1 + \epsilon)), \epsilon \right) - (\ln S)^2 \exp\left(-\frac{n}{4}\right) - \frac{\epsilon^2}{d_1^2} \cdot \sup_{x \in [d_1 - \epsilon, d_1 + \epsilon]} \ln^2(ex). \quad (101)$$

In light of Lemma 14, it suffices to consider $R_P(S, n, H, \epsilon)$ to give a lower bound of $R(S, n, H)$. Denote

$$\chi \triangleq \mathbb{E}_{\mu_1^S} H(P) - \mathbb{E}_{\mu_0^S} H(P) = 2\eta M E_L[-\ln x]_{[\eta, 1]} \cdot S = \frac{2d_1 H}{\ln S} \cdot E_L[-\ln x]_{[\eta, 1]} \succeq \frac{H}{\ln S} \cdot \ln \left(\frac{S \ln S}{nH \ln n} \right), \quad (102)$$

and

$$E_i \triangleq \mathcal{M}_S(\epsilon, H) \cap \left\{ P : |H(P) - \mathbb{E}_{\mu_i^S} H(P)| \leq \frac{\chi}{4} \right\}, \quad i = 0, 1. \quad (103)$$

Denote by π_i the conditional distribution defined as

$$\pi_i(A) = \frac{\mu_i^S(E_i \cap A)}{\mu_i^S(E_i)}, \quad i = 0, 1. \quad (104)$$

Now consider π_0, π_1 as two priors. By setting

$$\zeta = \mathbb{E}_{\mu_0^S} H(P) + \frac{\chi}{2}, \quad s = \frac{\chi}{4}, \quad \epsilon = \frac{1}{\ln n}, \quad (105)$$

we have $\beta_0 = \beta_1 = 0$ in Lemma 11. Applying union bound yields that

$$\mu_i^S[(E_i)^c] \leq \mu_i^S \left[\left| \sum_{j=1}^S p_j - d_1 \right| > \epsilon \right] + \mu_i^S \left[|H(P) - \mathbb{E}_{\mu_i^S} H(P)| > \frac{\chi}{4} \right] + \mu_i^S[H(P) > H] \quad (106)$$

and the Chebychev inequality tells us that

$$\mu_i^S \left[\left| \sum_{j=1}^S p_j - d_1 \right| > \epsilon \right] \leq \frac{1}{\epsilon^2} \sum_{j=1}^S \text{Var}_{\mu_i^S}(p_j) \leq \frac{SM^2}{\epsilon^2} \asymp \frac{S(\ln n)^4}{n^2} \preceq \frac{(\ln n)^4}{n^{\frac{1}{2}}} \rightarrow 0 \quad (107)$$

$$\mu_i^S \left[|H(P) - \mathbb{E}_{\mu_i^S} H(P)| > \frac{\chi}{4} \right] \leq \frac{16}{\chi^2} \sum_{j=1}^S \text{Var}_{\mu_i^S}(-p_j \ln p_j) \leq \frac{16S(M \ln M)^2}{\chi^2} \preceq \frac{S(\ln S)^2(\ln n)^4}{n^2} \preceq \frac{(\ln n)^6}{n^{\frac{1}{2}}} \rightarrow 0 \quad (108)$$

where we have used our assumption that $S \preceq n^{\frac{3}{2}}$. For bounding $\mu_i^S[H(P) > H]$, we first remark that for $d_1 \leq e^{-1}$,

$$\mathbb{E}_{\mu_i^S} H(P) \leq -d_1 \ln d_1 + (S-1) \int -t \ln t \mu_i(dt) \quad (109)$$

$$\leq -d_1 \ln d_1 - S \ln(\eta M) \int t \mu_i(dt) \quad (110)$$

$$= -d_1 \ln d_1 + \frac{d_1 H}{\ln S} \ln \left(\frac{S \ln S}{d_1 H} \right) \quad (111)$$

$$= -d_1 \ln d_1 + d_1 H - \frac{d_1 H}{\ln S} \ln \left(\frac{d_1 H}{\ln S} \right) \quad (112)$$

$$\leq d_1 H - 2d_1 \ln d_1 \quad (113)$$

hence, for d_1 sufficiently small, say, $d_1 \leq \min\{\frac{1}{4}, f^{-1}(\min\{\frac{H_0}{8}, \frac{1}{e}\})\}$, where $f(x) = -x \ln x$ is defined in $[0, e^{-1}]$ and $f^{-1}(\cdot)$ denotes the inverse function of $f(\cdot)$, we have

$$\mathbb{E}_{\mu_i^S} H(P) \leq d_1 H - 2d_1 \ln d_1 \leq \frac{H}{4} + 2 \cdot \min \left\{ \frac{H_0}{8}, \frac{1}{e} \right\} \leq \frac{H_0 + H}{4} \leq \frac{H}{2}. \quad (114)$$

Hence, similar to (108), we have

$$\mu_i^S[H(P) > H] \leq \mu_i^S \left[|H(P) - \mathbb{E}_{\mu_i^S} H(P)| > \frac{H}{2} \right] \leq \frac{S(M \ln M)^2}{(H/2)^2} \preceq \frac{S(\ln n)^4}{n^2} \preceq \frac{(\ln n)^4}{n^{\frac{1}{2}}} \rightarrow 0. \quad (115)$$

Denote by F_i, G_i the marginal probability under prior π_i and μ_i^S , respectively, for all $i = 0, 1$. In light of (106), (107), (108)

and (115), we have

$$V(F_i, G_i) \leq \mu_i^S[(E_i)^c] \rightarrow 0. \quad (116)$$

Moreover, by setting

$$d_1 = \min \left\{ \frac{1}{4}, f^{-1} \left(\min \left\{ \frac{H_0}{8}, \frac{1}{e} \right\} \right), \frac{1}{d_2^2 D_0} \right\}, \quad d_2 = 10e \quad (117)$$

it was shown in [11, Lem. 11] that

$$V(G_0, G_1) \leq \frac{S}{n^6} \preceq \frac{1}{n^{\frac{9}{2}}} \rightarrow 0. \quad (118)$$

Hence, the total variational distance is then upper bounded by

$$V(F_0, F_1) \leq V(F_0, G_0) + V(G_0, G_1) + V(G_1, F_1) \rightarrow 0 \quad (119)$$

where we have used the triangle inequality of the total variation distance. The idea of converting approximate priors μ_i^S into priors π_i via conditioning comes from Wu and Yang [15].

Now it follows from Lemma 11 and Markov's inequality that

$$R_P(S, n, H, \epsilon) \geq s^2 \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(\epsilon, H)} \mathbb{P}(|\hat{H} - H(P)| \geq s) \succeq \chi^2 \succeq \left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n} \right) \right]^2 \quad (120)$$

and the desired result follows directly from Lemma 14. Hence we have obtained the desired lower bound in the case $S \preceq n^{\frac{3}{2}}$.

For general $S \succeq n^{\frac{3}{2}}$, the non-decreasing property of $R(n, S, H)$ with respect to S shows that

$$R(n, S, H) \geq R(n, n^{\frac{3}{2}}, H) \succeq \left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n} \right) \right]^2 \Big|_{S=n^{\frac{3}{2}}} \asymp H^2 \quad (121)$$

which exactly equals to the desired lower bound $\left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n} \right) \right]^2$.

VII. FUTURE WORK

This paper studies the adaptive estimation framework to strengthen the optimality properties of the approximation theoretic entropy estimator proposed in Jiao et al. [11]. We remark that the techniques in this paper are by no means constrained to entropy, and we believe analogous results are also true for the estimators of $F_\alpha(P) = \sum_{i=1}^S p_i^\alpha$ in [11]. Furthermore, we find the fact that the sample size enlargement effect still holds in the adaptive estimation setting very intriguing, and we believe there is a larger picture surrounding this theme to be explored.

VIII. ACKNOWLEDGMENTS

The authors would like to express their most sincere gratitude to Dany Leviatan for valuable advice on the literature of approximation theory, in particular, for suggesting the result in Lemma 16.

APPENDIX A AUXILIARY LEMMAS

The following lemma characterizes the performance of the best uniform approximation polynomial for $-x \ln x, x \in [0, 1]$.

Lemma 15. Denote by $\sum_{k=0}^K g_{K,k} x^k$ the K -th order best uniform approximation polynomial for $-x \ln x, x \in [0, 1]$, then for $p_K(x) = \sum_{k=1}^K g_{K,k} x^k$, we have the norm bound

$$\sup_{x \in [0, 1]} |p_K(x) - (-x \ln x)| \leq \frac{D_n}{K^2} \quad (122)$$

where $D_n > 0$ is a universal constant for the norm bound. In fact, the following inequality holds:

$$\limsup_{K \rightarrow \infty} K^2 \cdot \sup_{x \in [0, 1]} |p_K(x) - (-x \ln x)| \leq \nu_1(2) \approx 0.453, \quad (123)$$

where the function $\nu_1(p)$ is was introduced by Ibragimov [20] as the following limit for p positive even integer and m positive integer

$$\lim_{n \rightarrow \infty} \frac{n^p}{(\ln n)^{m-1}} E_n[|x|^p \ln^m |x|]_{[-1, 1]} = \nu_1(p). \quad (124)$$

Furthermore, we also have the pointwise bound: there exists a universal constant $D_p > 0$ such that for any $C \geq 1$,

$$|p_K(x) - 2 \ln K \cdot x| \leq D_p C x, \quad \forall x \in \left[0, \frac{C}{K^2}\right]. \quad (125)$$

Lemma 16. [21, Thm. 8.4.8] There exists some universal constant $M > 0$ such that for any order- n polynomial $p(x)$ in $[0, 1]$, we have

$$\sup_{x \in [0, 1]} |p(x)| \leq M \cdot \sup_{x \in [n^{-2}, 1-n^{-2}]} |p(x)|. \quad (126)$$

The following lemma gives some tails bounds for Poisson and Binomial random variables.

Lemma 17. [22, Exercise 4.7] If $X \sim \text{Poi}(\lambda)$, or $X \sim \text{B}(n, \frac{\lambda}{n})$, then for any $\delta > 0$, we have

$$\mathbb{P}(X \geq (1 + \delta)\lambda) \leq \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}}\right)^\lambda \quad (127)$$

$$\mathbb{P}(X \leq (1 - \delta)\lambda) \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}}\right)^\lambda \leq e^{-\delta^2 \lambda / 2}. \quad (128)$$

APPENDIX B PROOF OF LEMMAS

A. Proof of Lemma 3

Denote $H(P) = \sum_{i=1}^S -p_i \ln p_i$ by H , we construct the lagrangian:

$$\mathcal{L} = \sum_{i=1}^S p_i (\ln p_i)^2 + \lambda \left(\sum_{i=1}^S -p_i \ln p_i - H \right) + \mu \left(\sum_{i=1}^S p_i - 1 \right). \quad (129)$$

By taking the derivative with respect to p_i , we obtain that

$$\frac{\partial \mathcal{L}}{\partial p_i} = (\ln p_i)^2 + 2 \ln p_i - \lambda(1 + \ln p_i) + \mu \quad (130)$$

is a quadratic form of $\ln p_i$, so the equation $\frac{\partial \mathcal{L}}{\partial p_i} = 0$ has at most two solutions.

Hence, we conclude that components of the maximum achieving distribution can only take two values $p_i \in \{q_1, q_2\}$, and suppose q_1 appears m times. Then our objective function becomes

$$\sum_{i=1}^S p_i (\ln p_i)^2 = \left(\sum_{i=1}^S p_i (\ln p_i)^2 \right) \left(\sum_{i=1}^S p_i \right) \quad (131)$$

$$= \left(\sum_{i=1}^S -p_i \ln p_i \right)^2 + \sum_{1 \leq i < j \leq S} p_i p_j (\ln p_i - \ln p_j)^2 \quad (132)$$

$$= H^2 + m(S - m)q_1 q_2 (\ln q_1 - \ln q_2)^2. \quad (133)$$

We distinguish the analysis into two cases.

1) *Case I:* If $\min\{q_1, q_2\} \geq \frac{1}{S^2}$, we have $-\ln p_i \leq 2 \ln S$ for all i . Hence,

$$\sum_{i=1}^S p_i (\ln p_i)^2 \leq 2 \ln S \cdot \sum_{i=1}^S -p_i \ln p_i = 2H \ln S. \quad (134)$$

2) *Case II:* If one of q_1 and q_2 is smaller than $\frac{1}{S^2}$, without loss of generality we can assume that $q_1 < \frac{1}{S^2}$. Then

$$\sum_{i=1}^S p_i (\ln p_i)^2 = H^2 + m(S - m)q_1 q_2 (\ln q_1 - \ln q_2)^2 \quad (135)$$

$$\leq H^2 + m(S - m)q_1 q_2 (\ln q_1)^2 \quad (136)$$

$$\leq H^2 + S q_1 (\ln q_1)^2 \quad (137)$$

$$\leq H^2 + S \cdot \frac{1}{S^2} \left(\ln \frac{1}{S^2} \right)^2 \quad (138)$$

$$= H^2 + \frac{4(\ln S)^2}{S} \quad (139)$$

where we have used the inequalities $m \leq S, (S - m)q_2 \leq 1$ and the monotonically increasing property of $x(\ln x)^2$ for $x \in [0, e^{-1}]$. Then the lemma is proved by noticing that $H \leq \ln S$.

B. Proof of Lemma 4

The lower bound follows directly from the concavity of $-x \ln x, 0 \leq x \leq 1$. For the upper bound,

$$p \ln n - \mathbb{E}[-\hat{p} \ln \hat{p}] = \mathbb{E}[\hat{p} \ln(n\hat{p})] \quad (140)$$

$$= \sum_{k=1}^n \frac{k \ln k}{n} \cdot \left(\mathbb{P}\left(\hat{p} \geq \frac{k}{n}\right) - \mathbb{P}\left(\hat{p} \geq \frac{k+1}{n}\right) \right) \quad (141)$$

$$= \sum_{k=1}^c \frac{k \ln k}{n} \mathbb{P}\left(\hat{p} = \frac{k}{n}\right) + \frac{c \ln c}{n} \mathbb{P}\left(\hat{p} = \frac{c}{n}\right) + \sum_{k=c+1}^n \frac{k \ln k - (k-1) \ln(k-1)}{n} \mathbb{P}\left(\hat{p} \geq \frac{k}{n}\right) \quad (142)$$

$$\leq \sum_{k=1}^c \frac{k \ln c}{n} \mathbb{P}\left(\hat{p} = \frac{k}{n}\right) + \frac{c \ln c}{n} \mathbb{P}\left(\hat{p} = \frac{c}{n}\right) + \sum_{k=c+1}^n \frac{k \ln k - (k-1) \ln(k-1)}{n} \mathbb{P}\left(\hat{p} \geq \frac{k}{n}\right) \quad (143)$$

$$\leq p \ln c + \frac{c \ln c}{n} \mathbb{P}\left(\hat{p} = \frac{c}{n}\right) + \sum_{k=c+1}^n \frac{k \ln k - (k-1) \ln(k-1)}{n} \mathbb{P}\left(\hat{p} \geq \frac{k}{n}\right). \quad (144)$$

The Chernoff bound yields

$$\mathbb{P}(n\hat{p} \geq (1+\delta)np) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^{np} \leq \left(\frac{e^{1+\delta}}{(1+\delta)^{1+\delta}} \right)^{np} = \left(\frac{e}{1+\delta} \right)^{(1+\delta)np}, \quad (145)$$

hence for any integer k , we have

$$\mathbb{P}\left(\hat{p} \geq \frac{k}{n}\right) = \mathbb{P}(n\hat{p} \geq k) \leq \left(\frac{enp}{k} \right)^k. \quad (146)$$

Then the desired result follows directly from the fact that

$$k \ln k - (k-1) \ln(k-1) = \ln k + (k-1) \ln \left(1 + \frac{1}{k-1} \right) \leq \ln k + 1. \quad (147)$$

C. Proof of Lemma 9

For the bias, it is straightforward to see that for $nX \sim \text{Poi}(np)$, we have

$$\mathbb{E}S_{K,H}(X) + p \ln p = \sum_{k=1}^K r_{K,H}(4\Delta)^{-k+1} p^k + p \ln p \quad (148)$$

$$= 4\Delta \left[p_K \left(\frac{p}{4\Delta} \right) - \ln(4\Delta) \cdot \frac{p}{4\Delta} \right] + p \ln p \quad (149)$$

$$= 4\Delta \left[p_K \left(\frac{p}{4\Delta} \right) - 2 \ln K \cdot \frac{p}{4\Delta} \right] + p \ln \left(\frac{c_2^2 n \ln n}{4c_1} \right) + p \ln p \quad (150)$$

where $p_K(x) \triangleq \sum_{k=1}^K g_{K,H} x^k$ is the best approximating polynomial appearing in Lemma 15. Since $\frac{p}{4\Delta} \leq \frac{1}{K^2}$, Lemma 15 asserts that

$$\left| p_K \left(\frac{p}{4\Delta} \right) - 2 \ln K \cdot \frac{p}{4\Delta} \right| \leq D_p \cdot \frac{p}{4\Delta} \quad (151)$$

and we conclude that

$$|\mathbb{E}S_{K,H}(X) + p \ln p| \leq -p \ln(pn \ln n) + (D_p + \ln(4c_1/c_2^2)) p. \quad (152)$$

The proof for the second part is similar to [11, Lem. 5].

D. Proof of Lemma 13

By defining

$$f_N(x) = -\ln \left(\frac{1+x}{2} + \frac{1-x}{2N} \right), \quad -1 \leq x \leq 1 \quad (153)$$

we have $E_L[f_N]_{[-1,1]} = E_L[-\ln x]_{[N^{-1},1]}$. Let $\Delta_L(x) = \frac{\sqrt{1-x^2}}{L} + \frac{1}{L^2}$ and define the following modulus of continuity for f :

$$\tau_1(f, \Delta_L) \triangleq \sup \{|f(x) - f(y)| : x, y \in [-1, 1], |x - y| \leq \Delta_L(x)\} \quad (154)$$

We have the following lemma.

Lemma 18. *There are an upper bound and a lower bound for $\tau_1(f_N, \Delta_L)$:*

$$\ln\left(\frac{N}{2L^2}\right) \leq \tau_1(f_N, \Delta_L) \leq \ln\left(\frac{2N}{L^2}\right), \quad \forall L \leq \frac{\sqrt{N}}{10} \quad (155)$$

Proof: The upper bound is shown in [15, Lem. 4]. For the lower bound, denote by $x_L \in [-1, 1]$ the solution to the equation $x_L - \Delta_L(x_L) = -1$, we have the following closed-form formula:

$$x_L = \frac{L^2 - L^4 + \sqrt{-3L^2 + L^4}}{L^2 + L^4} \geq -1 + \frac{1}{L^2}. \quad (156)$$

Hence, by definition, we have

$$\tau_1(f_N, \Delta_L) \geq |f_N(x_L) - f_N(-1)| = \ln\left(\frac{x_L + 1}{2}N + \frac{1 - x_L}{2}\right) \geq \ln\left(\frac{x_L + 1}{2}N\right) \geq \ln\left(\frac{N}{2L^2}\right). \quad (157)$$

■

The relationship between $\tau_1(f_N, \Delta_L)$ and $E_L[f_N]_{[-1,1]}$ was shown in [23, Thm. 3.13, Thm. 3.14] that there exist two universal constants $M_1, M_2 > 0$ such that

$$E_n[f_N]_{[-1,1]} \leq M_1 \tau_1(f_N, \Delta_n) \quad (158)$$

$$\frac{1}{n} \sum_{k=0}^n E_k[f_N]_{[-1,1]} \geq M_2 \tau_1(f_N, \Delta_n) \quad (159)$$

Applying (158) and (159) and setting the approximation order to be DL with constant $D > 1$ to be specified later, then given $N = (10D)^2 M \geq (10DL)^2$, the non-increasing property of $E_n[f_N]_{[-1,1]}$ with respect to n yields

$$E_L[f_N]_{[-1,1]} \geq \frac{1}{DL - L} \sum_{n=L+1}^{DL} E_n[f_N]_{[-1,1]} \quad (160)$$

$$\geq \frac{1}{DL} \left(\sum_{n=0}^{DL} E_n[f_N]_{[-1,1]} - E_0[f_N]_{[-1,1]} - \sum_{n=1}^L E_n[f_N]_{[-1,1]} \right) \quad (161)$$

$$\geq M_2 \tau_1(f_N, \Delta_{DL}) - \frac{\ln N}{DL} - \frac{M_1}{DL} \sum_{n=1}^L \tau_1(f_N, \Delta_n) \quad (162)$$

$$\geq M_2 \ln\left(\frac{N}{2(DL)^2}\right) - \frac{\ln N}{DL} - \frac{M_1}{DL} \sum_{n=1}^L \ln\left(\frac{2N}{n^2}\right) \quad (163)$$

$$\geq M_2 \ln\left(\frac{N}{2(DL)^2}\right) - \frac{\ln N}{DL} - \frac{M_1}{DL} \int_1^L \ln\left(\frac{2N}{x^2}\right) dx \quad (164)$$

$$\geq M_2 \ln\left(\frac{50K}{L^2}\right) - \frac{\ln K + 2 \ln(10D)}{DL} - \frac{M_1}{D} \ln\left(\frac{200e^2 D^2 K}{L^2}\right). \quad (165)$$

Hence, there exists a sufficiently large constant $D > 0$ such that

$$E_L[-\ln x]_{[(100D^2K)^{-1},1]} = E_L[f_N]_{[-1,1]} \geq \ln\left(\frac{K}{L^2}\right) \quad (166)$$

and this lemma is proved by setting $D_0 = \max\{100D^2, 1\}$.

E. Proof of Lemma 14

Fix $\delta > 0$. Let $\hat{H}(\mathbf{Z})$ be a near-minimax estimator of $H(P)$ under the Multinomial model. The estimator $\hat{H}(\mathbf{Z})$ obtains the number of samples n from observation \mathbf{Z} . By definition, we have

$$\sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_{\text{Multinomial}} |\hat{H}(\mathbf{Z}) - H(P)|^2 < R(S, n, H) + \delta, \quad (167)$$

where $R(S, n, H)$ is the minimax L_2 risk under the Multinomial model. Note that for any vector $P \in \mathcal{M}_S(\epsilon, H)$ (P is not necessarily a probability distribution), we have

$$H\left(\frac{P}{\sum_{i=1}^S p_i}\right) = \frac{H(P)}{\sum_{i=1}^S p_i} + \ln\left(\sum_{i=1}^S p_i\right) \leq \frac{H(P)}{d_1 - \epsilon} + \ln(d_1 + \epsilon) \quad (168)$$

where by definition we have $\left|\sum_{i=1}^S p_i - d_1\right| \leq \epsilon$. Hence, given $P \in \mathcal{M}_S(\epsilon, H)$, let $\mathbf{Z} = [Z_1, \dots, Z_S]^T$ with $Z_i \sim \text{Poi}(np_i)$ and let $n' = \sum_{i=1}^S Z_i \sim \text{Poi}(n \sum_{i=1}^S p_i)$, (168) suggests to use the estimator $d_1 \left(\hat{H}(\mathbf{Z}) - \ln d_1\right)$ to estimate $H(P)$. Note that

$$d_1 \left(\hat{H}(\mathbf{Z}) - \ln d_1\right) - H(P) = d_1 \left(\hat{H}(\mathbf{Z}) - H\left(\frac{P}{\sum_{i=1}^S p_i}\right)\right) + \left(\sum_{i=1}^S p_i\right) \ln\left(\sum_{i=1}^S p_i\right) - d_1 \ln d_1 \quad (169)$$

the triangle inequality gives (define $A = \sup_{x \in [d_1 - \epsilon, d_1 + \epsilon]} \ln^2(ex)$)

$$\frac{1}{2} \mathbb{E}_P \left| d_1 \left(\hat{H}(\mathbf{Z}) - \ln d_1\right) - H(P) \right|^2 \leq d_1^2 \mathbb{E}_P \left| \hat{H}(\mathbf{Z}) - H\left(\frac{P}{\sum_{i=1}^S p_i}\right) \right|^2 + \left| \left(\sum_{i=1}^S p_i\right) \ln\left(\sum_{i=1}^S p_i\right) - d_1 \ln d_1 \right|^2 \quad (170)$$

$$\leq d_1^2 \sum_{m=0}^{\infty} \mathbb{E}_P \left[\left| \hat{H}(\mathbf{Z}) - H\left(\frac{P}{\sum_{i=1}^S p_i}\right) \right|^2 \middle| n' = m \right] \mathbb{P}(n' = m) + \epsilon^2 A \quad (171)$$

$$\leq d_1^2 \sum_{m=0}^{\infty} R\left(S, m, \frac{H}{d_1 - \epsilon} + \ln(d_1 + \epsilon)\right) \mathbb{P}(n' = m) + \delta + \epsilon^2 A \quad (172)$$

$$\leq d_1^2 R\left(S, \frac{d_1 n}{2}, \frac{H}{d_1 - \epsilon} + \ln(d_1 + \epsilon)\right) \mathbb{P}(n' \geq \frac{d_1 n}{2}) + (d_1 \ln S)^2 \mathbb{P}(n' \leq \frac{d_1 n}{2}) + \delta + \epsilon^2 A \quad (173)$$

$$\leq d_1^2 R\left(S, \frac{d_1 n}{2}, \frac{H}{d_1 - \epsilon} + \ln(d_1 + \epsilon)\right) + (d_1 \ln S)^2 \exp\left(-\frac{d_1 n}{8}\right) + \delta + \epsilon^2 A, \quad (174)$$

where we have used the fact that conditioned on $n' = m$, $\mathbf{Z} \sim \text{Multinomial}(m, \frac{P}{\sum_{i=1}^S p_i})$, and $R(S, n, H) \leq (\sup_{P \in \mathcal{M}_S} H(P))^2 = (\ln S)^2$. Moreover, the last step follows from Lemma 17. The proof is completed by the arbitrariness of δ and Lemma 1.

F. Proof of Lemma 15

It has been shown in [11, Lemma 18] that

$$\lim_{K \rightarrow \infty} K^2 \cdot \sup_{x \in [0, 1]} \left| \sum_{k=0}^K g_{K,k} x^k - (-x \ln x) \right| = \frac{\nu_1(2)}{2}, \quad (175)$$

then plugging in $x = 0$ yields

$$\limsup_{K \rightarrow \infty} K^2 \cdot |g_{K,0}| \leq \frac{\nu_1(2)}{2}. \quad (176)$$

Hence, it follows from the triangle inequality that

$$\limsup_{K \rightarrow \infty} K^2 \cdot \sup_{x \in [0, 1]} \left| \sum_{k=1}^K g_{K,k} x^k - (-x \ln x) \right| \leq \frac{\nu_1(2)}{2} + \frac{\nu_1(2)}{2} = \nu_1(2), \quad (177)$$

which completes the proof of the norm bound.

For the pointwise bound, [21, Thm. 7.3.1] asserts that there exists a universal positive constant M_1 such that

$$\sup_{x \in [0, 1]} |(\varphi(x))^2 p_K''(x)| \leq M_1 K^2 \omega_\varphi^2(-x \ln x, K^{-1}), \quad (178)$$

where $\varphi(x) = \sqrt{x(1-x)}$, and $\omega_\varphi^2(f, t)$ is the second-order Ditzian-Totik modulus of smoothness [21] defined by

$$\omega_\varphi^2(f, t) \triangleq \sup \left\{ \left| f(u) + f(v) - 2f\left(\frac{u+v}{2}\right) \right| : u, v \in [0, 1], |u-v| \leq 2t\varphi\left(\frac{u+v}{2}\right) \right\}. \quad (179)$$

Direct computation yields

$$\omega_\varphi^2(-x \ln x, t) = \frac{2t^2 \ln 2}{1 + t^2}, \quad (180)$$

we have

$$\sup_{x \in [0,1]} |x(1-x)p_K''(x)| \leq 2M_1 \ln 2. \quad (181)$$

According to Lemma 16, since $p_K''(x)$ is a polynomial with order $K-2 < 2K$, there exists some positive constant M_2 such that

$$\sup_{x \in [0,1]} |p_K''(x)| \leq M_2 \sup_{x \in [(2K)^{-2}, 1-(2K)^{-2}]} |p_K''(x)| \quad (182)$$

$$\leq \frac{M_2(2K)^4}{(2K)^2 - 1} \sup_{x \in [(2K)^{-2}, 1-(2K)^{-2}]} |x(1-x)p_K''(x)| \quad (183)$$

$$\leq \frac{16M_2K^4}{4K^2 - 1} \sup_{x \in [0,1]} |x(1-x)p_K''(x)| \quad (184)$$

$$\leq \frac{32M_1M_2K^4 \ln 2}{4K^2 - 1} \quad (185)$$

$$\leq 16M_1M_2K^2 \ln 2, \quad (186)$$

hence for any $x, y \in [0, C/K^2]$, we have

$$|p_K'(x) - p_K'(y)| \leq \int_{\min\{x,y\}}^{\max\{x,y\}} |p_K''(t)| dt \leq 16M_1M_2 \ln 2 \cdot K^2 |x - y| \leq 16M_1M_2C \ln 2. \quad (187)$$

As a result, we know that for any $C \geq 1$ and $x \in [0, C/K^2]$,

$$16M_1M_2C \ln 2 \geq \frac{K^2}{C} \int_0^{\frac{C}{K^2}} |p_K'(x) - p_K'(t)| dt \quad (188)$$

$$\geq \left| p_K'(x) - \frac{K^2}{C} \int_0^{\frac{C}{K^2}} p_K'(t) dt \right| \quad (189)$$

$$= \left| p_K'(x) - \frac{K^2}{C} p_K \left(\frac{C}{K^2} \right) \right| \quad (190)$$

$$\geq \left| p_K'(x) - \frac{K^2}{C} \left(-\frac{C}{K^2} \ln \frac{C}{K^2} \right) \right| - \frac{K^2}{C} \left| p_K \left(\frac{C}{K^2} \right) - \left(-\frac{C}{K^2} \ln \frac{C}{K^2} \right) \right| \quad (191)$$

$$\geq |p_K'(x) - 2 \ln K| - \ln C - K^2 \sup_{t \in [0,1]} |p_K(t) - (-t \ln t)| \quad (192)$$

$$\geq |p_K'(x) - 2 \ln K| - \ln C - D_n, \quad (193)$$

where D_n is the coefficient of the norm bound in (122). Hence, the universal positive constant $D_p \triangleq 16M_1M_2 \ln 2 + 1 + D_n$ satisfies

$$|p_K'(x) - 2 \ln K| \leq D_p C, \quad \forall x \in \left[0, \frac{C}{K^2} \right], \quad (194)$$

and it follows that

$$|p_K(x) - 2 \ln K \cdot x| \leq \int_0^x |p_K'(t) - 2 \ln K| dt \leq \int_0^x D_p C dt = D_p C x. \quad (195)$$

REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.
- [2] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.
- [3] J. Hájek, "A characterization of limiting distributions of regular estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 4, pp. 323–330, 1970.
- [4] —, "Local asymptotic minimax and admissibility in estimation," in *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1972, pp. 175–194.
- [5] A. J. Wyner and D. Foster, "On the lower limits of entropy estimation," *IEEE Transactions on Information Theory*, submitted for publication, 2003.
- [6] B. Efron and R. Thisted, "Estimating the number of unseen species: How many words did shakespeare know?" *Biometrika*, vol. 63, no. 3, pp. pp. 435–447, 1976. [Online]. Available: <http://www.jstor.org/stable/2335721>

- [7] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM, 2009, pp. 49–62.
- [8] M. J. Berry, D. K. Warland, and M. Meister, "The structure and precision of retinal spike trains," *Proceedings of the National Academy of Sciences*, vol. 94, no. 10, pp. 5411–5416, 1997.
- [9] Z. F. Mainen and T. J. Sejnowski, "Reliability of spike timing in neocortical neurons," *Science*, vol. 268, no. 5216, pp. 1503–1506, 1995.
- [10] R. R. d. R. van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek, "Reproducibility and variability in neural spike trains," *Science*, vol. 275, no. 5307, pp. 1805–1808, 1997.
- [11] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distribution," *submitted to IEEE Trans. Inf. Theory*, 2014.
- [12] G. Valiant and P. Valiant, "Estimating the unseen: an $n/\log n$ -sample estimator for entropy and support size, shown optimal via new CLTs," in *Proceedings of the 43rd annual ACM symposium on Theory of computing*. ACM, 2011, pp. 685–694.
- [13] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
- [14] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Non-asymptotic theory for the plug-in rule in functional estimation," *arXiv preprint arXiv:1406.6959*, 2014.
- [15] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *available on arXiv*, 2014.
- [16] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Beyond maximum likelihood: from theory to practice," *arXiv preprint arXiv:1409.7458*, 2014.
- [17] T. T. Cai *et al.*, "Minimax and adaptive inference in nonparametric function estimation," *Statistical Science*, vol. 27, no. 1, pp. 31–50, 2012.
- [18] A. B. Tsybakov, "Aggregation and high-dimensional statistics," *Lecture notes for the course given at the École d'été de Probabilités in Saint-Flour*, URL http://www.crest.fr/ckfinder/userfiles/files/Pageperso/ATsybakov/Lecture_notes_SFfour.pdf, vol. 16, p. 20, 2013.
- [19] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.
- [20] I. I. Ibragimov, "Sur la valeur asymptotique de la meilleure approximation d'une fonction ayant un point singulier réel," *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, vol. 10, no. 5, pp. 429–460, 1946.
- [21] Z. Ditzian and V. Totik, *Moduli of smoothness*. Springer, 1987.
- [22] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.
- [23] P. P. Petrushev and V. A. Popov, *Rational approximation of real functions*. Cambridge University Press, 2011.